SUMMARY

This thesis focuses on studying the task of text-to-image generation, which aims at yielding photo-realistic and semantically consistent pictures on the basis of natural-language descriptions. Text-to-image synthesis has gathered growing attention in the research fields of computer vision and natural-language processing, since it plays an increasingly significant role in promoting the development of real artificial intelligence and satisfying the human appreciation and need for visual information and pictures. “What I cannot create, I do not understand”, said the famous physicist Richard Feynman. In spirit of this quote, we make the assumption that a machine will understand visual content if it is able to faithfully produce images based on their linguistic (textual) description. Thus, conducting the research in the domain of synthetic image generation from a text description may help a machine make sense of what they perceive. On the other hand, generating high-quality images according to textual input probes has a variety of potential applications, such as refinement of textual queries in internet, furniture design, art creation, data augmentation for training image classifiers, visually checking the output of captioning algorithms, photo editing according to textual descriptions.

There has been remarkable progress in this direction with the advances in a conditional generative adversarial network (cGAN). However, a cGAN-based text-to-image synthesis model still faces multiple challenges in the generation of visually appealing and semantically consistent samples as well as its applicability in practice: (1) the difficulty of directly approximating the high-resolution text-conditioned image distribution using a single generator/discriminator pair, (2) the frequent occurrences of the lack-of-diversity issue, (3) the high prior probability of non-successful patterns to be synthesized for a given random noise probe and (4) the poor explainability of a conditional text-to-image GAN framework are among the key challenges. To address these challenges, this thesis performs extensive studies on designing novel architectures, increasing the probability of drawing successful samples and improving the explainability of conditional text-to-image GAN models from the perspectives of both the latent space and the linguistic space. This dissertation can be summarized into seven chapters, as follows:

Chapter 1 provides a brief general introduction to the research of image synthesis based on linguistic (textual) descriptions. Furthermore, the research questions and significant contributions are explained.

Chapter 2 aims at producing perceptually plausible pictures that match with given natural-language descriptions, only employing a single generator/dis-
criminater pair consisting of a single network. To this end, we propose the Dual-Attention Generative-Adversarial Network (DTGAN) for text-to-image synthesis. Specifically, DTGAN introduces channel-aware and pixel-aware attention modules that can guide the generator to focus more on text-relevant channels and pixels by computing the attention weights between the global sentence vector and two aforementioned factors. More importantly, DTGAN spreads the attention over many, even all layers of the generator networks. This allows for an influence of the text over features at various hierarchical levels in the pipeline architecture, from crude early features to abstract late features. Afterwards, Conditional Adaptive Instance-Layer Normalization (CAdaILN) is developed to enable the linguistic cues from the sentence embedding to flexibly manipulate the amount of change in shape and texture, further improving visual-semantic representation and helping stabilize the training. Subsequently, a new type of visual loss is utilized to enhance the image resolution by ensuring vivid shape and perceptually uniform color distributions of generated images. The above four novel components contribute to a promising single-stage text-to-image synthesis framework. The results on benchmark data sets demonstrate the superiority of our proposed method compared to the state-of-the-art models with a multi-stage architecture.

Chapter 3 intends to deal with the lack-of-diversity issue present in the current single-stage text-to-image generation models that yield similar outputs given a single textual sequence and different injected random noise. To tackle this problem, we provide an efficient and effective single-stage framework (DiverGAN) to produce diverse, photo-realistic and semantically related images according to a single natural-language description and different latent vectors. DiverGAN develops two new word-level attention modules, i.e., a channel-attention module and a pixel-attention module, which model the importance of each word in the given sentence while allowing the network to assign larger weights to the significant channels and pixels semantically aligning with the salient words. With the help of these two word-related modulation modules, DiverGAN is able to effectively disentangle the attributes of the text description and accurately control the regions of synthetic images. After that, a dual-residual structure is designed to preserve more original visual features while allowing for deeper networks, resulting in faster convergence speed and more vivid details. More importantly, we propose to plug a fully-connected layer into the pipeline to address the lack-of-diversity problem, since we observe that a dense layer will remarkably enhance the generative capability of the network, balancing the trade-off between a low-dimensional random latent code contributing to variants and modulation modules that use high-dimensional and textual contexts to strengthen feature maps. Inserting a linear layer after the second residual block achieves the best variety and quality. Both qualitative and quantitative results on benchmark data sets suggest the effectiveness of DiverGAN for realizing diversity, without harming quality and semantic consistency.
Chapter 4 aims to ensure that generated samples are believable, realistic or natural. For this, we explore the semantic relationship between a realistic image and an inadequate sample in the space of generated images by linearly interpolating a successful starting-point latent code and an unsuccessful end-point latent vector. We discover that the first part of interpolation results is usually realistic but the final part is not plausible when conducting the linear interpolation between a Good latent code and a Bad latent vector. More importantly, the visual appearance and the semantic properties of interpolation samples do not always vary smoothly along with latent codes. We therefore make the assumption that there is a non-linear boundary separating high-resolution images from inadequate samples in the space of generated samples. Based on this assumption, we constructed two novel data sets (i.e., the Good & Bad bird and face data sets) that consist of successful as well as unsuccessful generated samples and are selected via strict criteria. To effectively and efficiently acquire high-quality images by increasing the probability of generating Good latent codes, we use a dedicated Good/Bad classifier for generated images. It is based on a pretrained front end and fine-tuned on the basis of the proposed Good & Bad data set. The results on the designed DiverGAN generator pre-trained on two benchmark data sets indicate that our classifier achieves a better than 98% accuracy in predicting Good/Bad classes for synthetic samples, which confirms the superiority of our presented techniques and data sets. Our data set is available on Zenodo1.

Chapter 5 is to examine the relationship between the latent control space and the obtained image variation in an unsupervised manner. To this end, we investigate the generation mechanism of a conditional text-to-image GAN model and further propose a novel approach for latent semantic discovery by conducting the independent component analysis (ICA) algorithm under an additional orthogonality constraint on the pretrained weight values of the generator. The captured directions are not only independent but also orthogonal. Furthermore, we develop a background-flattening loss (BFL) to improve the background appearance in an edited image. Also, we mathematically analyze the correspondences between Semantic Factorization (SeFa), GANSpace and regular PCA and show that they typically achieve almost the identical results when sampling enough data for GANSpace. The results on the recent single-stage text-to-image GAN models pre-trained on three benchmark data sets show that our presented method has the ability to not only derive various interpretable latent-space directions, but also provide a more precise control over the latent space than PCA.

Chapter 6 aims at taking a deep look into the linguistic space of a conditional text-to-image GAN model for an improved explainability. For this, we qualitatively analyze the roles played by linguistic embeddings in the synthetic-image semantic space through the linear interpolation between pairs of keywords. More specifically, we visualize the samples generated by linearly interpolating two

1 https://zenodo.org/record/6283798#.YhkN_ujMI2w
contrastive keywords, e.g., the color, the size, the length of the beak on the CUB bird data set and the background, the object, the action on the MSCOCO data set. Afterwards, we can observe the transition process from the initial image to the final sample. In addition, we extend a linear interpolation to a triangular interpolation for further analysis using a 2-simplex, i.e., triplet contrast, and better data augmentation. The results on the proposed DiverGAN generator trained on two benchmark data sets represent an improvement in explainability in the analyzed algorithm, which qualitatively suggests the effectiveness of our presented techniques.

Chapter 7 summarizes the main contributions of this thesis and provides answers to the research questions. In addition, several possible research directions that are associated with this thesis are discussed.